On-Policy Diffusion Reinforcement Learning Meets Off-Policy Quality Anchoring

Zunxu Liu[1], Zhaofan Qiu[2], Yazhen Xie[3]
Yingwei Pan[2], Ting Yao[2], and Tao Mei[2]

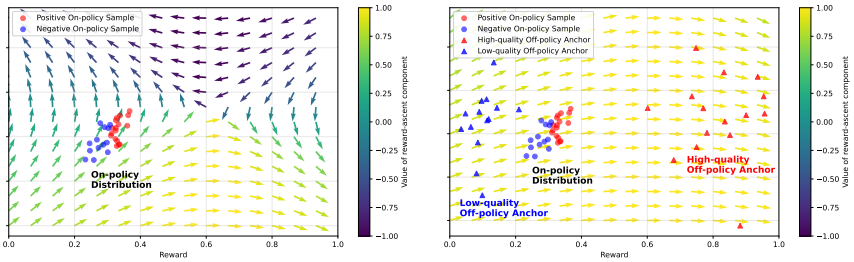
[1] University of Science and Technology of China, Hefei, China
[2] HiDream.ai Inc., Hefei, China
[3] Fudan University, Shanghai, China
irvingliu@mail.ustc.edu.cn, zhaofanqiu@gmail.com, yzxie24@m.fudan.edu.cn, {pandy,tiyao,tmei}@hidream.ai

Abstract. Reinforcement learning for diffusion models primarily relies on on-policy learning to optimize the model by using samples from its own current policy. Despite offering a direct optimization signal, the approach is fundamentally self-limiting and prone to reward overfitting and mode collapse due to its myopic guidance. In this paper, we propose a new framework, namely DiffusionCompass, that breaks such limitation by strategically integrating off-policy guidance. DiffusionCompass establishes a dual learning objective: the model continues to explore reward maximization on-policy, while simultaneously being anchored by off-policy examples that define a stable quality benchmark. That is achieved through a novel integration of several key techniques: an off-policy objective with distribution-aware filtering to adaptively emphasize relevant anchors, a perceptual anchored reward mechanism to mitigate reward hacking, and a reward renormalization strategy using off-policy samples for a stable quality baseline. The dual regimes complement each other: off-policy guidance prevents reward overfitting and collapse, and on-policy exploration ensures the model does not merely imitate a static dataset. Experiments across text-to-image/video generation demonstrate that DiffusionCompass achieves superior reward performance while maintaining high sample diversity, shaping a robust and effective path for fine-tuning generative models. Code is available at https://github.com/HiDream-ai/DiffusionCompass.

Keywords: Diffusion Models · Reinforcement Learning

1 Introduction

The advent of large-scale diffusion models has revolutionized image and video generation, enabling the synthesis of high-fidelity and diverse content from textual descriptions [3, 4, 6, 7, 11, 25, 31, 44, 46, 47, 54, 60, 61]. To further align these generative models with complex, non-differentiable objectives such as human preference or specific aesthetic criteria, reinforcement learning (RL) has emerged



(a) Only on-policy samples. **(b)** Both on-policy and off-policy samples.

Fig. 1: The optimization direction visualization, which represents the mean direction pointing towards the proximity of positive samples and away from negative samples.

as a powerful fine-tuning paradigm [12, 26, 35, 43, 52, 59]. The predominant approach in this domain is on-policy RL, where the model is optimized using reward signals derived from samples generated by its current policy [29, 42, 50, 52, 62]. Such self-referential learning mechanism provides a direct optimization signal. More specifically, by ranking and learning from its own outputs, the model can efficiently navigate the reward landscape within the neighborhood of its existing capabilities.

However, this on-policy paradigm is fundamentally self-limiting. By confining exploration strictly to the model's own narrow capability distribution, it is prone to two critical failure modes: (1) reward saturation, where the model quickly overfits to the locally optimal samples in its generation space and ceases to make further progress, and (2) mode collapse, a catastrophic drop in sample quality as the model collapses to generating only low-quality samples and fails to provide optimization direction towards high-quality samples. At its core, the model lacks an external, objective frame of reference for quality, rendering its understanding of performance self-referential purely defined by the rewards on its own restricted trajectory.

To break this cycle of myopic self-improvement, we introduce **DiffusionCompass**, a novel framework that remoulds standard on-policy RL through the integration of off-policy quality anchoring. Our key insight is that the shortcomings of on-policy exploration can be effectively mitigated by grounding the model's policy updates in a stable, comprehensive understanding of quality derived from a static dataset (i.e., off-policy samples). We visualize such motivation in Figure 1, where the optimization direction derived solely from on-policy samples is only locally accurate yet globally inconsistent due to its narrow perspective of the reward space (Figure 1a). In contrast, by incorporating both on-policy and off-policy samples, the off-policy anchors function analogously to a "compass," providing a more globally oriented optimization direction, thereby significantly accelerating the optimization process (Figure 1b). Technically, the on-policy component in our DiffusionCompass employs DiffusionNFT [62] to derive optimization direction from the reward ranking of sample groups generated online. The off-policy component complements this with a distribution-aware filtering mechanism for adaptively weighting off-policy anchors w.r.t. their proximity to the current policy, and a perceptually anchored reward for aligning on-policy

samples with high-quality references to combat reward hacking. Furthermore, a reward renormalization strategy provides a stable and globally-referenced baseline for on-policy learning. This synergy balances the system: off-policy anchors constrain the model from reward-hacking, while on-policy data prevents overfitting to static distributions.

Our primary contribution is a unified training framework, DiffusionCompass, that formally defines and implements this synergistic co-training through a holistic objective integrating these novel components. We demonstrate that the strategic integration of off-policy quality anchoring is not merely beneficial but essential for stable and effective diffusion RL. Extensive experiments on image and video generation tasks verify the effectiveness of our method, leading to both higher final reward and significantly improved convergence speed compared to strong on-policy baselines. By reconciling the exploratory drive of on-policy RL with the stable grounding of off-policy data, DiffusionCompass establishes a new paradigm for reliably fine-tuning generative models.

2 DiffusionCompass

This section presents the architecture of DiffusionCompass, our proposed framework for robust diffusion model post-training. We begin with a primer on the underlying flow matching models for visual generation. Next, we outline the standard on-policy paradigm, implemented via DiffusionNFT [62], and elaborate its inherent limitations. Our solution pivots to the strategic incorporation of off-policy data, which serves as a quality anchor to mitigate reward overfitting and prevent mode collapse. We conclude by formulating the holistic training objective that unifies on-policy exploration with off-policy guidance.

2.1 Preliminaries: Flow Matching

Flow matching has emerged as a foundational framework for high-performance visual generation [2,25,44,47], which transforms random noise into coherent data through a deterministic process. This is achieved by constructing a probability path from the data distribution $\mathbf{x}_0 \sim \mathcal{X}_0$ to a noise distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The Rectified Flow paradigm [30] defines this path with a linear trajectory:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\boldsymbol{\epsilon}. \tag{1}$$

The core learning target is the vector field $\mathbf{v} = \boldsymbol{\epsilon} - \mathbf{x}_0$, which defines the deterministic ordinary differential equation (ODE) for sampling. A conditional diffusion transformer (DiT) $\mathbf{v}_\theta(\mathbf{x}_t, t, c)$ is trained to approximate this field by minimizing the regression objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t} ||\mathbf{v} - \mathbf{v}_\theta(\mathbf{x}_t, c)||_2^2. \tag{2}$$

Here, c denotes the conditional text prompt. By accurately estimating the velocity field, the model can efficiently reverse the noising process to generate high-quality samples.

2.2 On-policy Self Exploration

While flow matching provides a powerful foundation for generative modeling through the regression loss in Eq.(2), this objective is primarily employed during pre-training to mimic the patterns of a pre-existing dataset. Consequently, models optimized solely for this objective often converge to a conservative equilibrium, producing technically coherent but aesthetically mediocre samples that merely recapitulate the training distribution. To steer the model beyond this plateau towards human-aligned objectives, we employ on-policy RL. This paradigm uses a reward signal to guide the model's self-exploration, pushing its policy to exploit regions of its capability that yield higher-scoring outputs.

However, directly applying classical on-policy methods like DPO [42, 52] or GRPO [29, 50] to diffusion models presents significant challenges. Their convergence is often slow, primarily due to the difficulty of constructing a rigorous probabilistic formulation for the multi-step sampling process. To circumvent this issue, we adopt DiffusionNFT [62] as our on-policy optimizer. Its key advantage lies in bypassing explicit probability modeling; instead, it directly leverages online-generated positive and negative samples to optimize the vector field in flow matching through a contrastive objective. This approach enables more efficient use of reward signals for faster convergence and more stable policy improvement.

Technically, we maintain a dual-model architecture: a trainable network $\mathbf{v}_\theta$ for policy updates and a snapshot network $\mathbf{v}_\theta^{\text{snap}}$, which is an Exponential Moving Average of the trainable network, providing stable sample generation. Initially, both networks are identically initialized. At each optimization step, we first use the snapshot network $\mathbf{v}_\theta^{\text{snap}}$ to generate a set of online samples, which are then evaluated by a reward model to obtain rewards. The exploring network $\mathbf{v}_\theta$ is subsequently optimized to align with the direction of policy improvement, as guided by the reward signal. Specifically, we first define the policy update as:

$$\Delta\mathbf{v}_\theta = \mathbf{v}_\theta - \mathbf{v}_\theta^{\text{snap}}. \tag{3}$$

Building upon the snapshot model as a starting point, we define a pair of vector fields that represent policy deviations along the estimated update direction:

$$\begin{aligned} \mathbf{v}_\theta^+ &:= \mathbf{v}_\theta^{\text{snap}} + \beta\Delta\mathbf{v}_\theta, \\ \mathbf{v}_\theta^- &:= \mathbf{v}_\theta^{\text{snap}} - \beta\Delta\mathbf{v}_\theta, \end{aligned} \tag{4}$$

where $\mathbf{v}_\theta^+$ and $\mathbf{v}_\theta^-$ denote the policies shifted towards and away from the update direction $\Delta\mathbf{v}_\theta$, respectively, and β is a scaling factor. The learning objective is to align $\mathbf{v}_\theta^+$ with the vector field of high-reward samples while aligning $\mathbf{v}_\theta^-$ with that of low-reward samples. This is formalized by the following loss function:

$$\mathcal{L}_{\text{ON}} = \mathbb{E}\left[p\left\|\mathbf{v}_\theta^+(\mathbf{x}_t, c) - \mathbf{v}\right\|_2^2 + (1-p)\left\|\mathbf{v}_\theta^-(\mathbf{x}_t, c) - \mathbf{v}\right\|_2^2\right]. \tag{5}$$

Here, $p \in [0, 1]$ represents the optimal probability, indicating how closely a sample $\mathbf{x}_0$ aligns with high reward. As proven in the original DiffusionNFT [62] work, optimizing this objective maximizes the expected reward.

In practice, since reward models output arbitrary scores rather than probabilities, we approximate p by processing a group of on-policy samples $\mathcal{X}_{\text{on}}$. Specifically, we normalize the rewards within this group by subtracting the mean and dividing by the standard deviation, clip the values to $[-1, 1]$, and then rescale them to the $[0, 1]$ range:

$$p \approx \frac{1}{2}\left(\text{clip}\left(\frac{r(\mathbf{x}_0, c) - \mu_{\mathcal{X}_{\text{on}}}}{\sigma_{\mathcal{X}_{\text{on}}}}, -1, 1\right) + 1\right), \tag{6}$$

where $\mu_{\mathcal{X}_{\text{on}}}$ and $\sigma_{\mathcal{X}_{\text{on}}}$ are the mean and standard deviation of rewards over $\mathcal{X}_{\text{on}}$. This approach follows the reward normalization technique used in GRPO [19].

2.3 Off-policy Guidance

The on-policy paradigm offers a direct, capacity-aware learning signal, yet its effectiveness is inherently constrained by a myopic, self-referential bias. Nevertheless, directly mixing off-policy samples with on-policy rollouts proves suboptimal, as off-policy samples inherently deviate from the current model's sampling distribution, which is the core assumption underpinning on-policy RL. To resolve this without disrupting on-policy dynamics, we introduce three complementary strategies that instead leverage off-policy data as quality anchors. First, we selectively incorporate off-policy samples aligned with the current on-policy distribution into the flow matching objective, weighting each by its distributional compatibility to filter out irrelevant examples. Second, to mitigate reward hacking, we compute semantic similarity between on-policy samples and their off-policy positive counterparts, up-weighting those that resemble high-quality anchors to steer optimization toward reliable high-reward regions. Third, we stabilize reward normalization by including off-policy rewards when estimating optimal probability, ensuring that a sample receives high weight only if it ranks favorably within both on-policy and off-policy sets for a given prompt. The remainder of this section details how these mechanisms are operationalized within our unified training framework.

(1) **Distribution-Aware Off-Policy Loss.** To mitigate the inherent limitations of on-policy methods, the off-policy objective should possess the following characteristics: 1. Straightforward: It bypasses complex policy optimization and directly guides the exploring network to align with high-quality outputs while diverging from low-quality ones. 2. Rational: Its rationality does not require off-policy samples to be drawn from the current policy. To achieve this, we first employ an offline model to generate N samples for each text prompt. The top $D\%$ of these samples with the highest rewards form the positive set $\mathcal{X}_{\text{off}}^{+}$, while the bottom $D\%$ constitute the negative set $\mathcal{X}_{\text{off}}^{-}$. This offline curation process provides a stable and consistent quality benchmark, which is more reliable and easier to learn from compared to the fluctuating on-policy samples.

For the positive samples, we directly minimize the distance between the predicted vector field $\mathbf{v}_\theta$ and the target vector field $\mathbf{v}$. However, for negative samples, simply maximizing this distance could lead to training instability, as the

distance is unbounded. Instead, inspired by concept-erasure techniques in diffusion models [13,29], we guide the network to associate negative samples with null prompts $\emptyset$. Specifically, we encourage the predicted vector field under an empty prompt $\mathbf{v}_\theta(\mathbf{x}_t, \emptyset)$ to approximate the vector field of negative samples. Leveraging classifier-free guidance, this approach effectively reduces the likelihood of generating low-quality samples during inference.

The aforementioned learning objective remains suboptimal, as it assigns equal guidance strength to all off-policy positive and negative samples, regardless of their relation to the current policy distribution. For positive samples, prioritizing high-quality instances far outside the current policy's support can cause large off-policy gradients and destabilize training. Instead, positive samples near the current policy distribution are more informative and should be emphasized. For negative samples, it is more meaningful to suppress those the policy is likely to generate than to penalize already unlikely ones. Therefore, we introduce a distribution-aware filtering mechanism. Specifically, for each off-policy sample, we introduce an empirically effective approximate distance measure d to quantify its distance from the current policy distribution:

$$d = \mathrm{sg}(\|\mathbf{v} - \mathbf{v}_\theta\|_2^2) \tag{7}$$

where sg denotes the stop-gradient operation. Here, d serves as a more efficient way to quantify the distance between an off-policy sample and the policy distribution. Subsequently, we introduce a soft-filtering weight w for both positive and negative off-policy samples $w = e^{-\alpha d}$, which adjusts the off-policy guidance strength in a distribution-aware manner, allowing for differential emphasis on samples based on their proximity to the current policy distribution. Finally, the off-policy loss is formulated as:

$$\mathcal{L}_{\mathrm{OFF}} = \mathbb{E}_{\mathbf{x}_0 \in \mathcal{X}^+_{\mathrm{off}}} w_{\mathbf{x}_0} \|\mathbf{v} - \mathbf{v}_\theta(\mathbf{x}_t, c)\|_2^2 + \mathbb{E}_{\mathbf{x}_0 \in \mathcal{X}^-_{\mathrm{off}}} w_{\mathbf{x}_0} \|\mathbf{v} - \mathbf{v}_\theta(\mathbf{x}_t, \emptyset)\|_2^2. \tag{8}$$

(2) **Perceptual Anchored Reward.** Reward models provide only a single, and often biased, scalar reward. Such bias can mislead the policy optimization process. When the model overfits to the specific weaknesses of the reward function instead of genuinely improving task performance, a phenomenon known as reward hacking may occur. In contrast, off-policy positive data offer greater flexibility: they can be manually curated to serve as a more comprehensive anchor signal, emphasizing not only high reward scores but also high-quality outputs. Therefore, we leverage off-policy positive data to encourage the policy model to produce high-scoring samples that are perceptually similar to these strong references, thus mitigating the risk of mode collapse. Specifically, we extract features from the corresponding off-policy data and compute their similarity with features extracted from the on-policy samples. The maximum similarity value is then used as an auxiliary reward to refine the original reward signal.

Specifically, for sample $\mathbf{x}_0$, we calculate perceptual anchored reward:

$$r_p(\mathbf{x}_0, t) = r(\mathbf{x}_0, t) \cdot \max_{\hat{\mathbf{x}} \in \mathcal{X}_{\mathrm{off}}} \mathrm{similarity}\big(\mathrm{f}_{\mathbf{x}_0}, \mathrm{f}_{\hat{\mathbf{x}}}\big), \tag{9}$$

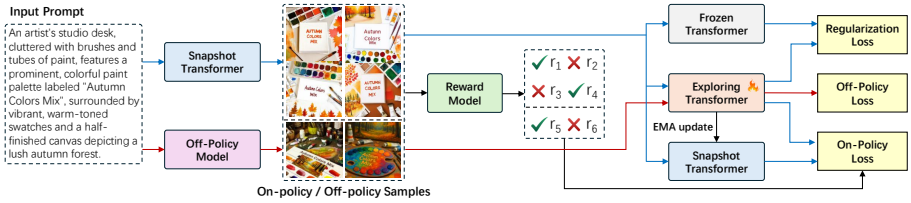


Fig. 2: The overall framework of DiffusionCompass. In each iteration: 1) The on-policy flow (**blue arrows**) uses the snapshot model to generate samples, which are rewarded and jointly normalized with off-policy samples to compute $\mathcal{L}_{\text{ON}}$ via reward renormalization and the NFT objective. 2) The off-policy flow (**red arrows**) retrieves pre-generated high- and low-quality samples to compute the anchoring loss $\mathcal{L}_{\text{OFF}}$. These losses, together with a KL regularizer $\mathcal{L}_{\text{KL}}$, form the final objective $\mathcal{L}$ to update the trainable model, whose weights are EMA-transferred to the snapshot model..

where $\mathrm{f}_{\mathbf{x}_0}$ and $\mathrm{f}_{\hat{\mathbf{x}}}$ denote the feature representations from the on-policy sample $\mathbf{x}_0$ and off-policy reference samples $\hat{\mathbf{x}}$, respectively. The features are extracted using DINOv2 [32] and averaged across frames on video data.

(3) **Reward Renormalization.** To establish a stable and globally referenced reward baseline, we integrate off-policy samples into the normalization process. Specifically, at each update step, we augment the on-policy sample group $\mathcal{X}_{\text{on}}$ with a curated batch of off-policy samples $\mathcal{X}_{\text{off}}$ sourced from external guidance models. The reward normalization for estimating the optimal probability p is then performed over this combined set:

$$p \approx \frac{1}{2}\left(\text{clip}\left(\frac{r(\mathbf{x}_0, c) - \mu_{\mathcal{X}_{\text{on}} \cup \mathcal{X}_{\text{off}}}}{\sigma_{\mathcal{X}_{\text{on}} \cup \mathcal{X}_{\text{off}}}}, -1, 1\right) + 1\right). \tag{10}$$

This simple yet effective reward renormalization technique confers two key advantages: it stabilizes the reward signal by providing a consistent quality baseline, which mitigates the variance from homogeneous on-policy groups, and it naturally steers the model towards higher-quality outputs from the outset by anchoring the reward distribution to include genuine positive examples.

2.4 Overall Framework

Figure 2 illustrates the overall framework of our proposed DiffusionCompass, where the on-policy and off-policy data flows are depicted with blue and red arrows, respectively. In each iteration, we first sample a prompt from the prompt set. The snapshot transformer is then used to generate a group of on-policy samples $\mathcal{X}_{\text{on}}$, while a corresponding set of off-policy samples $\mathcal{X}_{\text{off}}$ is retrieved from a pre-generated offline collection. All samples are evaluated by the reward model to obtain their reward signals. The overall training objective combines three key components: the on-policy loss $\mathcal{L}_{\text{ON}}$ that drives policy improvement, the off-policy loss $\mathcal{L}_{\text{OFF}}$ that provides quality anchoring, and a KL-divergence loss $\mathcal{L}_{\text{KL}}$ that constrains the exploring network from deviating too far from the frozen initial network, thereby ensuring training stability. The complete optimization objective of DiffusionCompass is formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{ON}} + \lambda_2 \mathcal{L}_{\text{OFF}} + \lambda_3 \mathcal{L}_{\text{KL}}, \tag{11}$$

where λ_1, λ_2, and λ_3 are trade-off parameters that balance the contributions of the respective components.

2.5 Speed-up by Dynamic Sampling

A significant bottleneck in standard on-policy fine-tuning for diffusion models is the substantial computational overhead of generating samples during training. Given the multi-step nature of the diffusion process, producing a single video can take several minutes, rendering iterative on-policy exploration prohibitively slow. To mitigate this, we propose a **Dynamic Sampling** strategy that drastically accelerates training by generating low-fidelity "snapshots" for on-policy evaluation, rather than full-quality samples.

Conventionally, such aggressive down-sampling in resolution and frame rate would severely degrade the training signal. However, within our framework, the high-quality, static off-policy corpus provides a stable and comprehensive anchor for visual quality and temporal coherence. This allows the on-policy component to focus on efficient exploration using these lightweight snapshots, without collapsing to low-quality modes. Specifically, we employ distinct sampling strategies for different reward types:

(1) **Space-related Reward.** For rewards evaluating spatial quality (e.g., aesthetics, fidelity), we generate on-policy samples using single-frame sampling. This provides a sufficient signal for spatial attributes while dramatically reducing sampling time and training overhead.

(2) **Time-related Reward.** For rewards evaluating temporal dynamics (e.g., motion coherence), we generate on-policy samples at a low spatial resolution and a high frame number. This strategy effectively strengthens the model's generative capacity at higher frame rates, improving attributes like background consistency and motion smoothness, but at a fraction of the computational cost.

3 Related Work and Discussion

Our work sits at the intersection of diffusion model fine-tuning and reinforcement learning. We contextualize DiffusionCompass by relating it to three key areas.

Fine-Tuning for Diffusion Models. A primary goal in generative modeling is to align pre-trained diffusion models with nuanced, often non-differentiable, objectives. Early approaches largely fall into two categories. Training-free methods [55], such as classifier-free guidance [13, 21, 37, 40] or direct noise optimization [41], adjust the sampling process without updating model weights. Alternatively, supervised fine-tuning methods, including continued training [16, 36] or direct reward regression [8, 10, 35], adapt the model parameters on curated datasets. However, these methods are fundamentally constrained by the quality and diversity of the static dataset, often leading to overfitting and a limited exploration of the reward landscape.

Reinforcement learning has emerged as a powerful paradigm to overcome these limitations, enabling direct optimization of complex reward functions. Dif-

fusionCompass follows this direction but critically addresses the core instability of existing on-policy RL methods by integrating a stable, off-policy data anchor.

Reinforcement Learning for LLMs. The success of RL in aligning Large Language Models (LLMs) with human preferences provides a valuable blueprint for generative models. Policy gradient algorithms, progressing from PPO [39] to more efficient methods like GRPO [19] and DAPO [57], have demonstrated significant improvements. A key insight emerging from recent LLM research is the inherent limitation of purely on-policy learning, which can trap the model within its own current policy distribution, leading to slow progress and mode collapse [48, 51, 58]. This has spurred the development of off-policy guidance techniques to improve sample training stability.

Our work is grounded in a parallel observation for diffusion models: exclusive reliance on on-policy data induces a myopic self-improvement loop. DiffusionCompass translates the conceptual solution of off-policy guidance from the LLM domain to the unique challenges of diffusion model fine-tuning, establishing a synergistic training objective that balances exploration with quality anchoring.

Reinforcement Learning for Diffusion Models. The application of RL to diffusion models is an active area of research. Existing methods can be broadly categorized into reward-weighted regression [15, 26], direct preference optimization [42, 52] and its variants [18, 28], and PPO-based policy gradients [1, 12]. A recent and influential trend involves adapting the GRPO framework from LLMs to diffusion models, often by reformulating the deterministic ODE sampler into a stochastic SDE process to define a policy [27,29,50]. While these methods show promise, their tight coupling with the SDE sampler can compromise sampling efficiency and final output quality. DiffusionNFT [62] addresses this by aligning the RL objective more closely with the standard flow-matching loss, leading to faster convergence. However, as an on-policy method, it remains susceptible to the fundamental issues of reward overfitting, training instability, and high computational cost for online sample generation, which are acutely magnified in resource-intensive domains like video generation.

DiffusionCompass directly builds upon and extends this line of work. We adopt the efficient on-policy learning but fundamentally augment it with a static off-policy corpus. This integration mitigates the instability and myopia of pure on-policy learning, prevents mode collapse, and, through dynamic sampling of on-policy data, dramatically reduces training costs without sacrificing final performance. Our work demonstrates that the strategic inclusion of off-policy guidance is a critical next step for stable and scalable diffusion RL.

4 Experiments

4.1 Task and Dataset.

Image Synthesis. We follow established protocols to assess the enhancement of image flow-matching models on three critical tasks: (1) **Compositional Image Generation**, benchmarked on GenEval [17]; (2) **Visual Text Rendering**,

measured via PaddleOCR [9] for OCR accuracy; (3) **Human Preference Alignment**, evaluated using the human preference metric PickScore [24]. As noted in the previous work [29], tasks (1) and (2) often lead to image quality degradation during fine-tuning, while task (3) frequently suffers from a loss of diversity. Our experiments systematically analyze these trade-offs to validate the robustness of our method.

Video Synthesis. For video generation, we employ the comprehensive VBench benchmark [22], which evaluates models across 16 distinct dimensions. These dimensions can be categorized into three groups: (1) **Spatial Quality** (e.g., aesthetics, object generation, spatial relationships); (2) **Temporal Dynamics** (e.g., object/background consistency, motion degree); (3) **Semantic Quality** (e.g., human action accuracy). To the best of our knowledge, there has been no prior work that directly leverages the full VBench metrics, a highly complex multi-reward framework, as reward models for post-training video generation systems. This presents a substantial challenge, which our proposed method is specifically designed to alleviate.

4.2 Implementation Details

Our DiffusionCompass is built upon the Diffusers library [34] in Python. Unless otherwise specified, our implementation follows the major design elements and experimental protocols of FlowGRPO and DiffusionNFT within the Diffusers library [34] to guarantee a fair comparison.

Architectures. For image generation, we use SD3.5-Medium at a resolution of 512×512 as the base model and Qwen-Image [47] as the off-policy model. For video generation, we build upon Wan2.1-1.3B at 480p resolution as the base model and Wan2.1-14B as off-policy model.

Training Configuration. For image tasks, we align with standard practices by fine-tuning the pre-trained model using LoRA with a rank $r = 32$ and scaling parameter $\alpha = 64$. The parameters of the Lora are optimized by AdamW optimizer with a fixed learning rate of 3×10^{-4}. We adopt the loss weights $\lambda_1 = 1$, $\lambda_2 = 0.1$, and $\lambda_3 = 1 \times 10^{-4}$. The online samples are generated with 10 sampling steps but evaluated with 40 steps for final assessment. For video tasks, the rollout sampling steps are set to 20, while evaluation remains at 40 steps.

4.3 Evaluation Metrics

Beyond the task-specific metrics (GenEval, OCR accuracy, PickScore, VBench), we introduce several general-purpose metrics to comprehensively quantify model performance, with a particular focus on generalization and the mitigation of reward hacking: **Aesthetic Score** [38]: Assesses the overall visual quality. **CLIP-Score** [20]: Measures the semantic alignment between text prompts and generated images. **Group SSIM** [45]: Estimates the structural similarity within a group of images generated from the same prompt; a higher value indicates lower diversity. **Vendi Score** [14]: Evaluates the diversity of the generated image distribution; a higher score indicates greater diversity.

Table 1: Overall performance comparisons across multiple benchmarks.The best scores are in **bold**, and the second-best scores are underlined.

(a) GenEval benchmark.

Model	Iter	Overall	Single Obj.	Two Obj.	Counting	Colors	Position	Attr. Bind
FLUX.1-Dev [2]	-	0.66	0.98	0.81	0.74	0.79	0.22	0.45
SD3.5-L [11]	-	0.71	0.98	0.89	0.73	0.83	0.34	0.47
Janus-Pro-7B [5]	-	0.80	0.99	0.89	0.59	0.90	0.79	0.66
GPT-4o [23]	-	0.84	0.99	0.92	0.85	0.92	0.75	0.61
Qwen-Image [47]	-	0.91	**1.00**	0.95	0.93	0.92	0.87	0.83
SD3.5-M [11]	-	0.63	0.98	0.78	0.50	0.81	0.24	0.52
FlowGRPO [29]	5K	0.95	**1.00**	**1.00**	0.95	0.91	0.94	0.87
DiffusionNFT [62]	1K	0.95	**1.00**	**1.00**	**0.97**	0.92	0.97	0.83
DiffusionCompass	1K	**0.97**	**1.00**	**1.00**	**0.97**	**0.93**	**0.99**	**0.91**

(b) OCR benchmark.

Model	GPU hours	OCR
SD3.5-M	-	0.55
FlowGRPO (with CFG)	600	0.915
DiffusionNFT with CFG	64	0.937
DiffusionNFT w/o CFG	64	0.970
DiffusionCompass with CFG	**64**	0.965
DiffusionCompass w/o CFG	**64**	**0.975**

(c) PickScore.

Model	PickScore
FlowGRPO	23.4
DiffusionNFT	23.7
DiffusionCompass	**23.8**

(d) Out-of-domain on OCR.

Model	In-domain	Image Qual.	Text Align.
	OCR	Aesthetic	ClipScore
DiffusionNFT	0.970	4.45	0.280
DiffusionCompass	**0.975**	**5.00**	**0.306**

(e) Out-of-domain on PickScore.

Model	In-domain	Aes.	Clip	Group SSIM ↓	Pixel Vendi	Incep. Vendi
	PickScore					
DiffusionNFT	23.7	**6.13**	0.270	0.324	1.46	2.01
DiffusionCompass	**23.8**	**6.13**	**0.285**	**0.262**	**1.87**	**2.45**

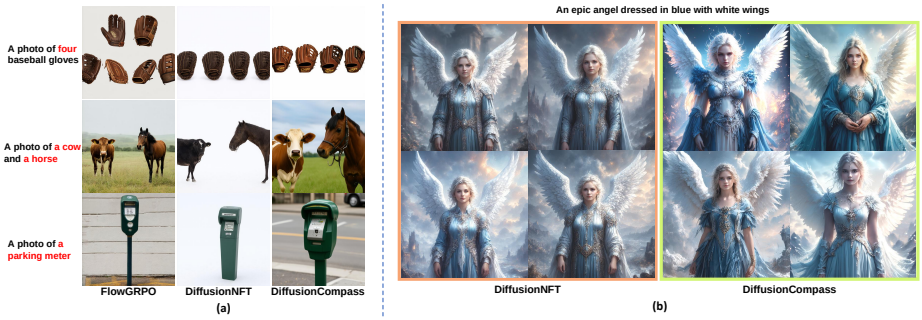


Fig. 3: Qualitative Comparisons on (a) GenEval and (b) PickScore benchmark.

4.4 Performance Comparisons and Analysis

Evaluation on Compositional Image Generation. As summarized in Table 1a, DiffusionCompass achieves the best overall performance on the GenEval compositional image generation benchmark. Specifically, it attains an overall score of 0.97, outperforming both FlowGRPO and DiffusionNFT (each scoring 0.95) while requiring only 1K training iterations, one-fifth of FlowGRPO's 5K. This result also surpasses our off-policy baseline Qwen-Image (0.91), underscoring the effectiveness of our hybrid on-policy and off-policy strategy. Visual comparisons in Figure 3 further illustrate these advantages. Unlike competing approaches that often sacrifice semantic alignment or fall prey to reward hacking (manifesting as repetitive patterns or monotonous backgrounds), DiffusionCompass maintains high image-text consistency without compromising aesthetic quality. This success indicates the necessity of off-policy guidance as a safeguard against over-optimization, thereby enhancing generalization in image synthesis.

Evaluation on Visual Text Rendering and Human Preference Alignment. As shown in Table 1b, our method demonstrates consistent advantages on the text rendering benchmark (OCR): DiffusionCompass achieves higher OCR reward than both FlowGRPO and DiffusionNFT, with and without CFG, while requiring only 64 GPU hours for training, significantly less than FlowGRPO's 600 GPU hours. Visual results in Figure 4 depict that DiffusionCompass gener-

Table 2: Performance comparisons with state-of-the-art text-to-video generation models on VBench benchmark. There are 16 dimensions in total, including Subject Consistency (SC), Background Consistency (BC), Temporal Flickering (TF), Motion Smoothness(MS), Dynamic Degree (DD), Aesthetic Quality (AQ), Imaging Quality (IQ), Object Class (OC), Multiple Objects (MO), Human Action(HA), Color (C), Spatial Relationship (SR), Scene (S), Appearance Style (AS), Temporal Style (TS), Overall Consistency (OC'). RP: Refined prompt proposed in Wan2.1 [44].

Methods	Iter	Total	SC	BC	TF	MS	DD	AQ	IQ	OC	MO	HA	C	SR	S	AS	TS	OC'
Open-Sora-2.0 [33]	-	81.71	**98.75**	98.00	99.40	**99.49**	20.74	64.33	65.62	94.50	77.72	95.40	85.98	76.18	52.71	22.98	**25.91**	27.57
CogVideoX-5B [53]	-	81.91	96.45	96.71	98.97	97.20	69.51	61.88	63.33	85.07	63.94	98.60	83.03	68.91	51.96	**24.98**	25.42	**27.65**
CogVideoX1.5-5B [53]	-	82.01	96.56	96.81	98.53	98.15	56.16	62.07	65.34	83.42	67.28	97.60	88.40	79.43	53.28	24.68	25.42	27.42
Hunyuan Video [25]	-	83.43	97.22	97.60	99.39	99.05	71.94	60.28	67.24	83.48	66.71	94.40	89.79	72.13	54.46	22.21	24.52	26.95
CausVid [56]	-	83.88	98.09	97.41	96.89	97.81	80.60	64.70	69.67	92.80	71.68	**99.80**	80.34	64.77	55.84	24.18	25.29	27.53
Wan2.1-1.3B [44]	-	78.90	93.58	95.95	99.19	96.88	79.86	59.06	63.61	77.81	48.06	77.50	78.66	70.41	20.06	20.44	22.60	23.06
Wan2.1-1.3B [44]+RP	-	81.23	93.68	97.20	99.24	98.24	59.72	63.88	64.45	90.98	71.49	87.00	79.09	73.99	46.22	20.42	23.76	25.65
Wan2.1-14B [44]+RP	-	83.69	97.52	**98.09**	**99.46**	98.30	64.46	66.07	69.43	86.28	69.58	95.40	88.59	75.39	45.75	22.64	23.19	25.91
DiffusionNFT(1.3B)+RP	320	83.52	92.29	97.37	99.15	98.31	72.22	65.10	69.11	95.97	84.45	92.00	89.05	81.42	49.71	20.42	23.11	25.89
ours (1.3B)	320	82.53	95.04	94.80	99.10	97.65	**91.67**	61.38	67.39	87.58	70.20	84.00	88.77	80.86	29.72	21.21	24.00	24.84
ours (1.3B)+RP	320	85.00	94.26	96.95	99.35	98.41	81.94	66.28	**70.71**	**96.84**	88.41	90.00	88.01	86.48	50.36	20.19	23.87	26.16
ours (14B)+RP	160	**87.16**	95.74	97.43	99.37	98.72	84.44	**68.82**	70.43	96.76	**93.13**	97.40	**93.58**	**91.78**	**66.74**	23.01	25.21	26.90

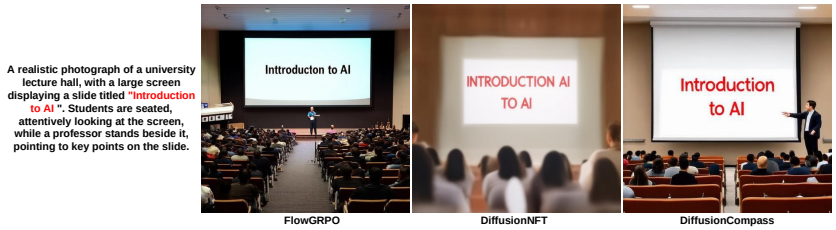


Fig. 4: Qualitative Comparisons on OCR benchmark.

ates more accurate and legible text renderings. A similar trend is observed for human preference alignment in Table 1c. While PickScore optimization appears saturated, DiffusionCompass (23.8) still outperforms both FlowGRPO (23.4) and DiffusionNFT (23.7). More importantly, qualitative results in Fig. 3 reveal a stark contrast in generalization: DiffusionNFT suffers from severe mode collapse, generating nearly identical images across random seeds due to reward over-optimization. In contrast, DiffusionCompass maintains superior diversity without compromising performance, indicating robustness against such collapse.

Evaluation on Out-of-domain Performance. Tables 1d and 1e illustrate that DiffusionCompass leads to a superior balance between reward maximization and generalization. In the OCR task, it surpasses DiffusionNFT in both image quality (5.00 vs. 4.45) and text alignment (0.306 vs. 0.280). Moreover, for the PickScore task, our method exhibits robust diversity preservation while maintaining high reward performance. Notably, DiffusionCompass significantly outperforms the baseline on diversity indicators, as evidenced by a lower Group SSIM (0.262 vs. 0.324) and improved Pixel Vendi scores (1.87 vs. 1.46). Such performance underscores its ability to circumvent the mode collapse that typically plagues baseline models.

Evaluation on Comprehensive Video Generation. As shown in Table 2, DiffusionCompass obtains steady performance improvement on the VBench benchmark throughout training. After post-training, the total score of Wan1.3B with

Fig. 5: Qualitative Comparisons on VBench benchmark.

Table 3: Ablation studies on OCR benchmark for (a) each designs in DiffusionCompass; (b) the mixed strategy; (c) sensitivity to λ_2; (d) sensitivity to α.

(a)

Reward Renorm	Anch. Reward	Anch. Loss	Dist. Filter	GPUh	OCR	Aes.	Clip
-	-	-	-	64	0.970	4.45	0.280
✓	-	-	-	56	0.969	4.48	0.282
✓	✓	-	-	56	0.963	**5.03**	**0.309**
✓	-	✓	-	64	0.974	4.66	0.298
✓	✓	✓	-	64	0.972	4.88	0.302
✓	✓	✓	✓	64	**0.975**	5.00	0.306

(b)

Model	OCR
SD3.5-M	0.550
SD3.5-M + Off-policy RL	0.836
SD3.5-M + On-policy RL	0.970
DiffusionCompass	**0.975**

(c)

λ_2	0.001	0.01	0.1	0.5
GenEval	0.96	**0.97**	**0.97**	0.92
OCR	0.970	**0.975**	**0.975**	0.959

(d)

α	1	10	100	1000
GenEval	0.94	**0.97**	**0.97**	0.94
OCR	0.959	**0.975**	**0.975**	0.961

refined prompts increases from 81.23 to 85.00, surpassing not only its 14B-parameter counterpart (83.69) but also other state-of-the-art models like CausVid and HunyuanVideo. This is particularly notable as VBench integrates 16 distinct evaluation dimensions, making direct improvement highly challenging. Certain metrics, such as dynamic degree, are especially prone to reward hacking, which might degrade video quality or cause training collapse. Nevertheless, our method enables consistent gains across most metrics, proving robustness under multi-reward optimization. In addition, DiffusionCompass with 14B parameters reaches 87.16, demonstrating the scalability of our approach. Qualitative results in Figure 5 confirm these improvements, showing enhanced text alignment, improved adherence to style instructions, naturalness of human actions, motion dynamics, and spatial consistency with the prompt compared to the base model.

4.5 Ablation Studies

Additional robustness and generalization. We further validate the robustness of DiffusionCompass from three aspects. (1) For the anchor source in Table 4a, replacing Qwen-Image with Flux only slightly changes aesthetic quality, while using SD3.5-M's own initial samples as anchors still achieves an OCR score of 0.973, close to 0.975 with Qwen-Image. This self-anchoring setting removes

Table 4: Additional robustness and generalization studies for (a) off-policy models; (b) OCR performance based on FlowGRPO; (c) additional evaluations on ImageReward (IR) and VideoOCR (VO) .

(a)

Anchor	OCR	Aes.
SD3.5-M	0.973	4.92
Flux	0.973	4.94
Qwen-Image	**0.975**	**5.00**

(b)

Method	OCR	Aes.	Clip
FlowGRPO	0.915	5.04	0.310
+ Ours	**0.939**	**5.08**	**0.320**

(c)

Method	IR	VO
DiffusionNFT	1.64	0.412
Ours	**1.70**	**0.433**

the possibility of simply distilling a stronger model, suggesting that the key factor is the relative quality contrast provided by off-policy anchors rather than the absolute strength of the anchor generator. (2) For the on-policy learner in Table 4b, we evaluate a stricter on-policy backbone, FlowGRPO, where samples are generated from the active policy rather than from an EMA snapshot as in DiffusionNFT. Incorporating our off-policy anchoring improves OCR from 0.915 to 0.939, while also increasing aesthetics from 5.04 to 5.08 and ClipScore from 0.310 to 0.320. These consistent gains indicate that DiffusionCompass is not a method-specific patch for DiffusionNFT, but a general mechanism that can complement different on-policy diffusion RL objectives. (3) For the evaluation scope in Table 4c, the improvements in ImageReward [49] from 1.64 to 1.70 and in VideoOCR from 0.412 to 0.433 show that the off-policy anchor helps both perceptual preference alignment and video text rendering. Overall, these studies support the view that off-policy quality anchoring serves as a reusable stabilizing signal: it is robust to anchor source, compatible with different on-policy learners, and beneficial across multiple reward dimensions.

Effect of designs of DiffusionCompass. We analyze the impact of each component in DiffusionCompass on the Visual Text Rendering task (Table 3a). Reward renormalization alone prevents training failures caused by reward homogenization, reducing training time while maintaining performance. Adding anchored reward steers the model towards high-quality samples, boosting out-of-domain performance. The anchoring loss further refines this effect by ensuring the model moves closer to high-quality samples, improving both in-domain and out-of-domain results. Finally, distribution filtering accounts for policy mismatch, delivering the best overall performance with the highest OCR (0.975), Aes. (5.00), and ClipScore (0.306). Each component contributes to better alignment with the optimal policy. We further evaluate hyperparameter sensitivity in Tables 3c and 3d. DiffusionCompass remains stable across practical ranges of the off-policy loss weight λ_2 and the distribution-filtering strength α, with the best OCR and GenEval scores obtained around $\lambda_2 \in \{0.01, 0.1\}$ and $\alpha \in \{10, 100\}$.

Effect of Mixed Policy Strategy. As shown in Table 3b, we compare DiffusionCompass against single-policy variants (pure on-policy or off-policy) on the OCR benchmark. While each individual policy improves reward, the mixed strategy in DiffusionCompass yields the highest performance, demonstrating the complementary benefits of combining both policies.

Convergence Speed with Off-Policy Anchors. Table 5a illustrates the effect of incorporating varying numbers of off-policy samples (n per prompt)

Table 5: Ablation studies on (a) convergence across OCR steps (n: off-policy samples per prompt); (b) dynamic sampling on VBench spatial relationship.

(a)

Method \Step	0	8	16	24	32	40	48	56
Only On-policy	0.412	0.484	0.640	0.832	0.938	0.947	0.980	0.976
+ Off-policy (n=2)	0.412	0.495	0.663	0.826	0.941	0.946	0.979	0.975
+ Off-policy (n=4)	0.412	0.723	0.900	0.930	0.968	0.952	0.977	**0.977**
+ Off-policy (n=6)	0.412	**0.795**	**0.957**	**0.973**	**0.986**	**0.983**	**0.983**	**0.977**

(b)

Model	GPU hours	Spatial relation.
Wan2.1-1.3B	-	73.99
DiffusionCompass w/o dynamic sampling	120	77.40
DiffusionCompass with dynamic sampling	**60**	**86.45**

on convergence speed. Using only two samples shows slight improvement over pure on-policy training, as such a small set provides insufficient signal for reliable quality assessment. In contrast, including four or more samples significantly accelerates convergence, confirming the value of off-policy guidance.

Role of Dynamic Sampling. Dynamic sampling is essential for efficient video generation. Ablation results in Table 5b on spatial relationship reward (VBench) indicate that removing this component markedly slows training speed and degrades performance under a fixed runtime budget.

Fairness Consideration. Given the significant computational disparity between on-policy and off-policy training, where each off-policy step requires approximately 10% the time of an on-policy iteration, we normalize iteration counts to ensure fair convergence comparisons. Specifically, we equate ten off-policy iterations to one on-policy iteration, and apply this conversion when reporting total iterations for our mixed-policy approach.

5 Conclusion

In this work, we proposed DiffusionCompass, a unified framework that strategically augments on-policy exploration with off-policy quality anchoring. Our method establishes a synergistic dual objective where the model simultaneously explores the reward landscape through its own generated samples and is grounded by a stable benchmark of pre-curated high-quality and low-quality examples. Extensive experiments on text-to-image and text-to-video generation tasks verify that DiffusionCompass achieves superior performance while showing higher convergence speed, demonstrating a more robust and effective path for alignment. Furthermore, the integration of dynamic sampling indicates the framework's practical advantage in significantly accelerating training without compromising final model quality. By providing a stable "compass" for navigation in the complex reward landscape, our work establishes a new paradigm for reliable and efficient generative model fine-tuning. The principles of DiffusionCompass are model-agnostic, showing strong potential for application across a wide range of generative tasks.

Acknowledgements. This work was supported by the Key Science & Technology Project of Anhui Province No. 202523o09050002.

References

1. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
2. BlackForest: Flux. `https://github.com/black-forest-labs/flux`, accessed 29 June 2026 (2024)
3. Cai, Q., Chen, J., Gao, C., Gong, Z., Li, Y., Pan, Y., Peng, Y., Qiu, Z., Yu, K., Zhang, Y., Ai, H., Bai, S., Chen, Y., Chen, Z., Gao, F., Guo, Y., Li, D., Shen, Z., Shi, L., Wang, J., Wang, S., Wang, Y., Zheng, R., Yao, T., Mei, T.: Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. arXiv preprint arXiv:2605.11061 (2026)
4. Chen, J., Long, F., An, J., Qiu, Z., Yao, T., Luo, J., Mei, T.: Ouroboros-diffusion: Exploring consistent content generation in tuning-free long video diffusion. In: AAAI (2025)
5. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
6. Chen, Z., Long, F., Qiu, Z., Yao, T., Zhou, W., Luo, J., Mei, T.: Aligning global semantics and local textures in generative video enhancement. In: ICCV (2025)
7. Chen, Z., Long, F., Qiu, Z., Yao, T., Zhou, W., Luo, J., Mei, T.: Tuning-free high-resolution video diffusion with spatial-temporal latent grouping. IEEE TMM (2025)
8. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. arXiv preprint arXiv:2309.17400 (2023)
9. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., et al.: Paddleocr 3.0 technical report. arXiv preprint arXiv:2507.05595 (2025)
10. Ding, Z., Jin, C., Liu, D., Zheng, H., Singh, K.K., Zhang, Q., Kang, Y., Lin, Z., Liu, Y.: Dollar: Few-step video generation via distillation and latent reward optimization. In: ICCV (2025)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
12. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In: NeurIPS (2023)
13. Frans, K., Park, S., Abbeel, P., Levine, S.: Diffusion guidance is a controllable policy improvement operator. arXiv preprint arXiv:2505.23458 (2025)
14. Friedman, D., Dieng, A.B.: The vendi score: A diversity evaluation metric for machine learning. arXiv preprint arXiv:2210.02410 (2022)
15. Furuta, H., Zen, H., Schuurmans, D., Faust, A., Matsuo, Y., Liang, P., Yang, S.: Improving dynamic object interactions in text-to-video generation with ai feedback. arXiv preprint arXiv:2412.02617 (2024)
16. Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., et al.: Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346 (2025)
17. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. In: NeurIPS (2023)
18. Gu, Y., Wang, Z., Yin, Y., Xie, Y., Zhou, M.: Diffusion-rpo: Aligning diffusion models through relative preference optimization. arXiv preprint arXiv:2406.06382 (2024)

19. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
20. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: EMNLP (2021)
21. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
22. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR (2024)
23. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
24. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: NeurIPS (2023)
25. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
26. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
27. Li, J., Cui, Y., Huang, T., Ma, Y., Fan, C., Yang, M., Zhong, Z.: Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. arXiv preprint arXiv:2507.21802 (2025)
28. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Li, J., Zheng, L.: Step-aware preference optimization: Aligning preference with denoising performance at each step. arXiv preprint arXiv:2406.04314 (2024)
29. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. In: NeurIPS (2025)
30. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)
31. Long, F., Qiu, Z., Yao, T., Mei, T.: Videostudio: Generating consistent-content and multi-scene videos. In: ECCV (2024)
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
33. Peng, X., Zheng, Z., Shen, C., Young, T., Guo, X., Wang, B., Xu, H., Liu, H., Jiang, M., Li, W., et al.: Open-sora 2.0: Training a commercial-level video generation model in $200 k. arXiv preprint arXiv:2503.09642 (2025)
34. Platen, P.V., et al.: Diffusers: State-Of-The-Art Diffusion Models. `https://github.com/huggingface/diffusers`, accessed 29 June 2026 (2022)
35. Prabhudesai, M., Goyal, A., Pathak, D., Fragkiadaki, K.: Aligning text-to-image diffusion models with reward backpropagation. arXiv preprint arXiv:2310.03739 (2023)
36. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
37. Sadat, S., Kansy, M., Hilliges, O., Weber, R.M.: No training, no problem: Rethinking classifier-free guidance for diffusion models. arXiv preprint arXiv:2407.02687 (2024)

38. Schuhmann., C.: Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, accessed 29 June 2026 (2022)
39. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
40. Shen, D., Song, G., Xue, Z., Wang, F.Y., Liu, Y.: Rethinking the spatial inconsistency in classifier-free diffusion guidance. In: CVPR (2024)
41. Tang, Z., Peng, J., Tang, J., Hong, M., Wang, F., Chang, T.H.: Tuning-free alignment of diffusion models with direct noise optimization. In: ICML 2024 Workshop on Structured Probabilistic Inference and Generative Modeling (2024)
42. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: CVPR (2024)
43. Wan, S., Chen, J., Cai, Q., Pan, Y., Yao, T., Mei, T.: VTON-VLLM: Aligning virtual try-on models with human preferences. In: NeurIPS (2025)
44. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
45. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP (2004)
46. Wu, A., Qiu, Z., Yao, T., Mei, T.: Ps-sr: Pseudo-single-step video super-resolution via speculative diffusion. In: CVPR (2026)
47. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
48. Wu, J., Liao, C., Feng, M., Zhang, S., Wen, Z., Luo, H., Yang, L., Xu, H., Tao, J.: Templaterl: Structured template-guided reinforcement learning for llm reasoning. arXiv preprint arXiv:2505.15692 (2025)
49. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Neurips (2023)
50. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
51. Yan, J., Li, Y., Hu, Z., Wang, Z., Cui, G., Qu, X., Cheng, Y., Zhang, Y.: Learning to reason under off-policy guidance. In: NeurIPS (2025)
52. Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., Li, X.: Using human feedback to fine-tune diffusion models without any reward model. In: CVPR (2024)
53. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
54. Yao, T., Li, Y., Pan, Y., Qiu, Z., Mei, T.: Denoising token prediction in masked autoregressive models. In: ICCV (2025)
55. Yeh, P.H., Lee, K.H., Chen, J.C.: Training-free diffusion model alignment with sampling demons. In: ICLR (2025)
56. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: CVPR (2025)
57. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)

58. Zhang, W., Xie, Y., Sun, Y., Chen, Y., Wang, G., Li, Y., Ding, B., Zhou, J.: On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. arXiv preprint arXiv:2508.11408 (2025)
59. Zhang, Y., Qiu, Z., Liu, Z., Pan, Y., Yao, T., Mei, T.: Evoid: Reinforced evolution for identity-preserving video generation. In: CVPR (2026)
60. Zhang, Z., Long, F., Pan, Y., Qiu, Z., Yao, T., Cao, Y., Mei, T.: Trip: Temporal residual learning with image noise prior for image-to-video diffusion models. In: CVPR (2024)
61. Zhang, Z., Long, F., Qiu, Z., Pan, Y., Liu, W., Yao, T., Mei, T.: Motionpro: A precise motion controller for image-to-video generation. In: CVPR (2025)
62. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., et al.: Diffusionnft: Online diffusion reinforcement with forward process. In: ICLR (2026)